データ粒子化による高速高精度な次世代マイニング技術の創出

宇野 毅明　（国立情報学研究所
＆総合研究大学院大学）

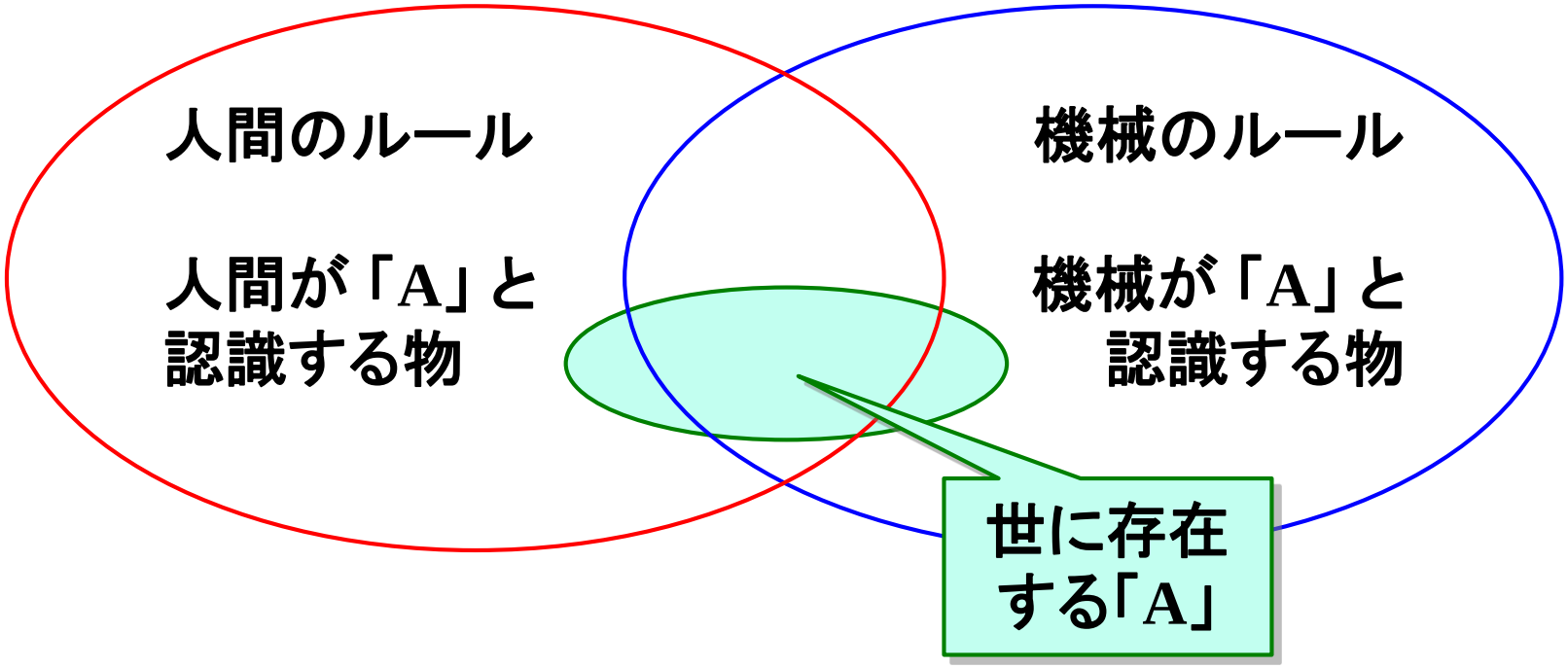
http://research.nii.ac.jp/~uno/index-j.html
e-mail: uno@nii.ac.jp

2017年12月25日　河原林ERATO Winter Festa Ep.3

解がわかりにくい

ディープラーニングなど、データマイニングや機械学習の出す解は、理解しにくかったり、大量でえらさがわからなかったり

機械が作りやすいルールだったり、機械が見付けやすい解を列挙しているのだから、そりゃ人間にはわかりにくいでしょう



わかやすくするには

マイニングと学習を人間にわかりやすくするなら、人間の認知プロセスに沿ったアプローチをとらないと

人間が物事を観察して、なんらかの概念を得るとき

　抽象化や具体化を行い、仮説を立て、それをデータで検証しという、抽象度合いや粒度の大きさを変えながらデータを見回して、考えていく

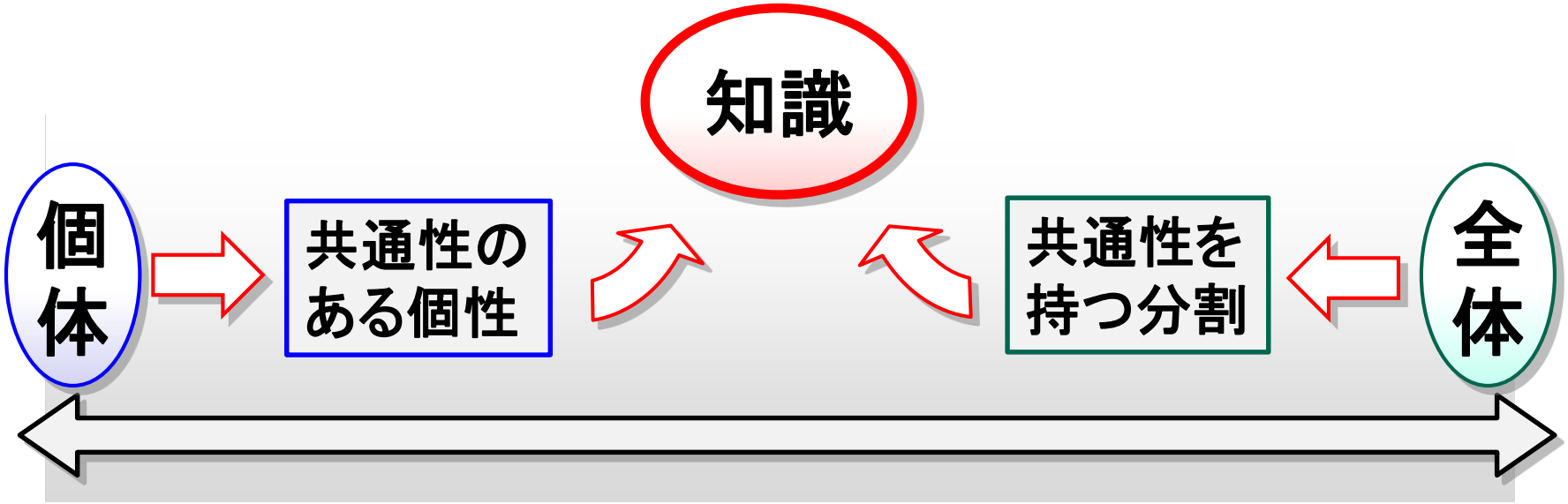
たとえば、このような過程に寄り添った解析手法を作れば良いだろう

人間の認知活動に沿った解析

- 人がデータを見るとき、アプローチは2つ

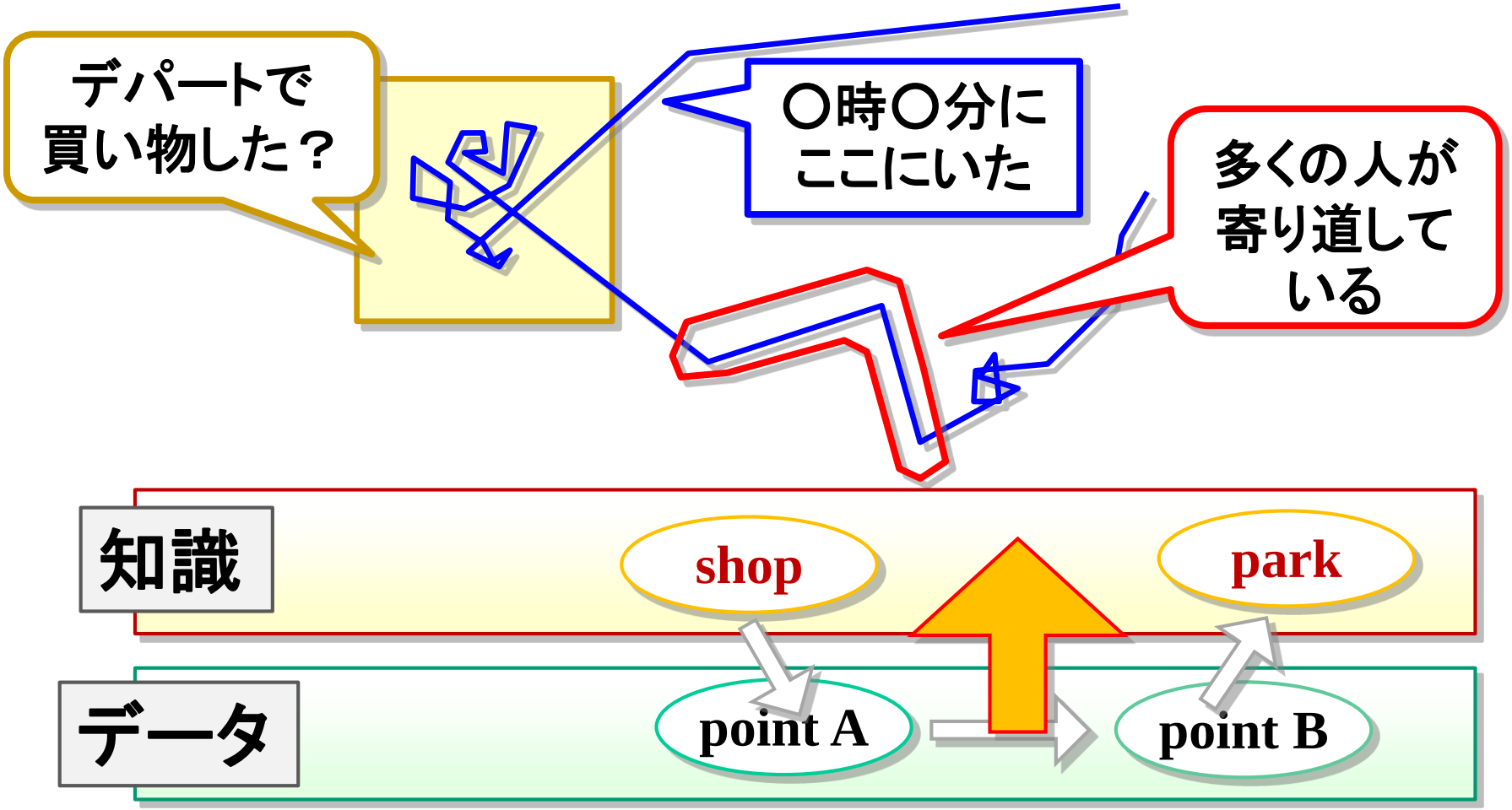
＋ 全体を眺め、いくつかに分割し、その組合せや関係性、違いをもとに知識を得る

＋ いくつかの個体をじっくり観察し、その個性の中から共通性を見つけ出し、その共通性とデータやシステムとの関係性から知識を得る



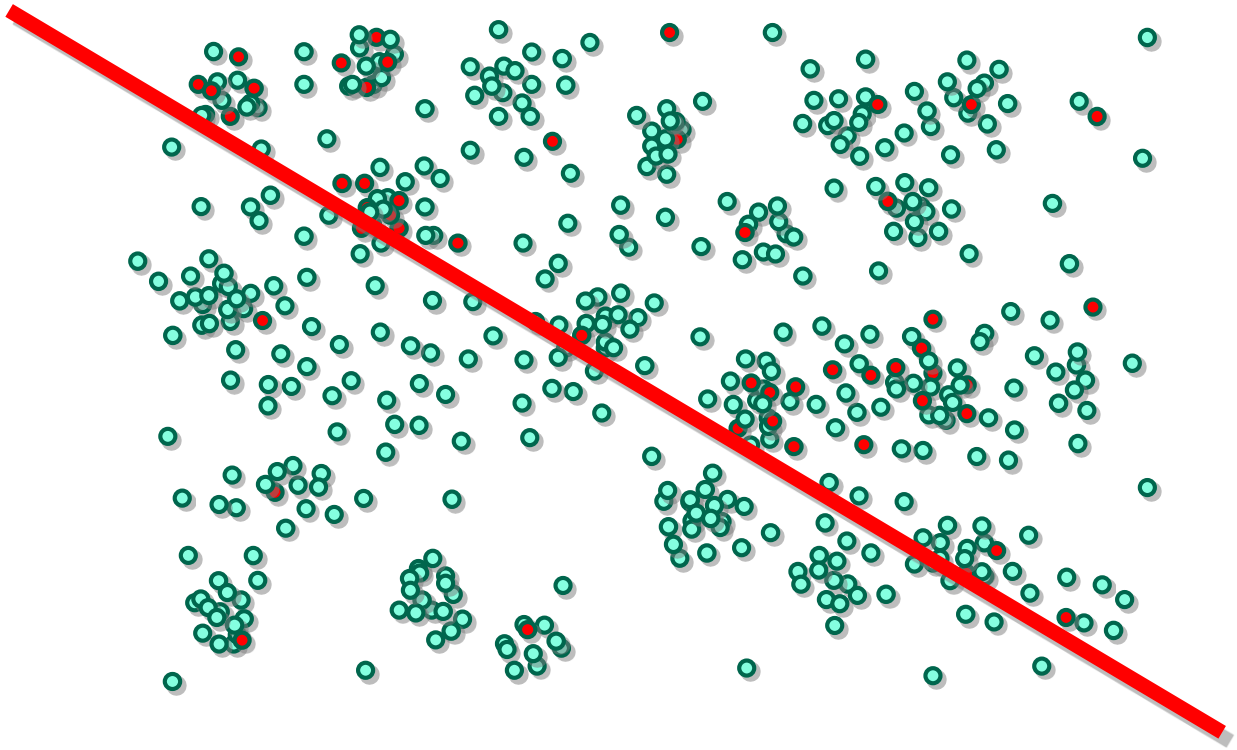
分かりにくいデータ

- 位置情報などのセンサデータ、非常に正確で密度も濃いが、だからと言って意味が取りやすいわけではない
- 抽象度の高い、意味の濃い部分（粒子）は、理解しやすい



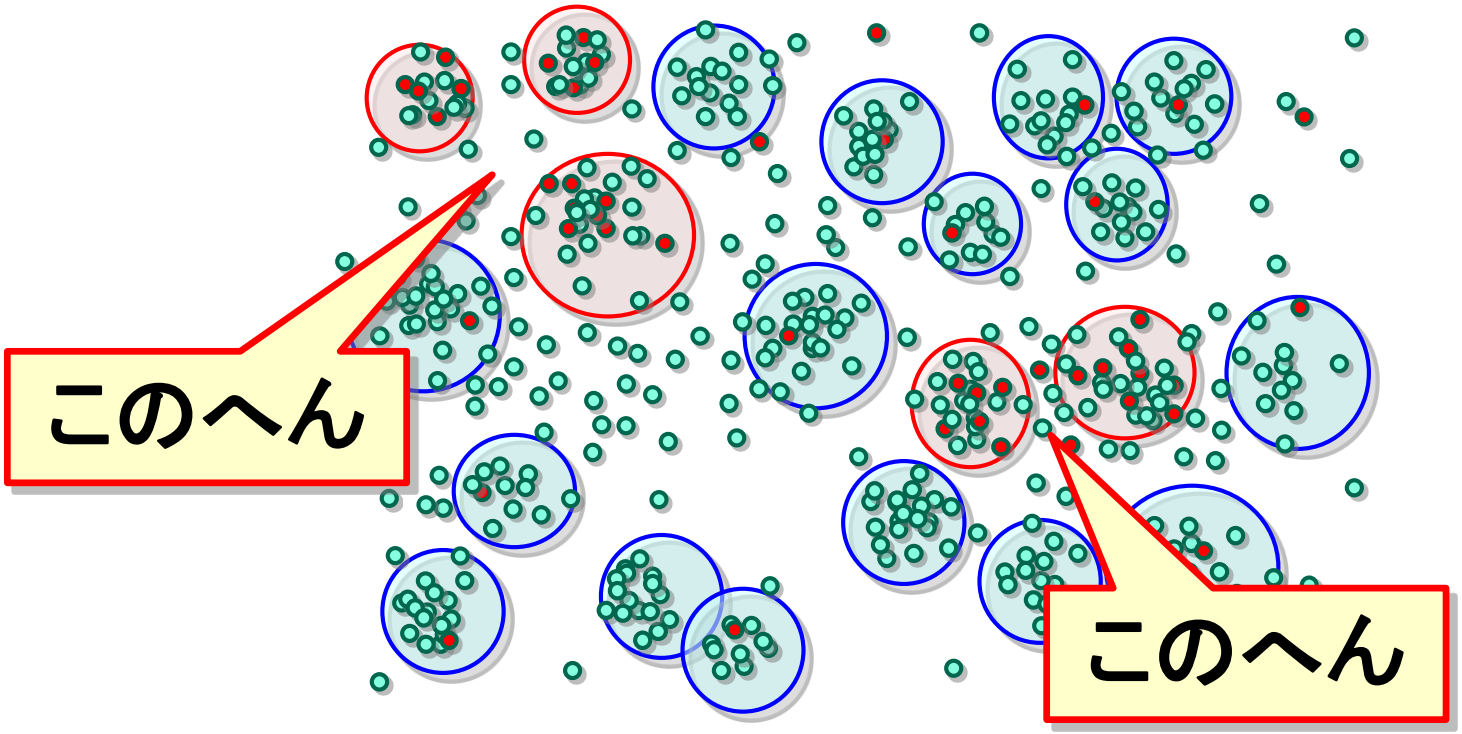
機械学習

- ○と×を分けるルールを、データから見つける手法
- 分け方／切り方により多くの方法がある
- 解（切り方）は、基本的に**「理解しがたい」**



粒子による学習

- 粒子を使って解を作ると、すっきりする

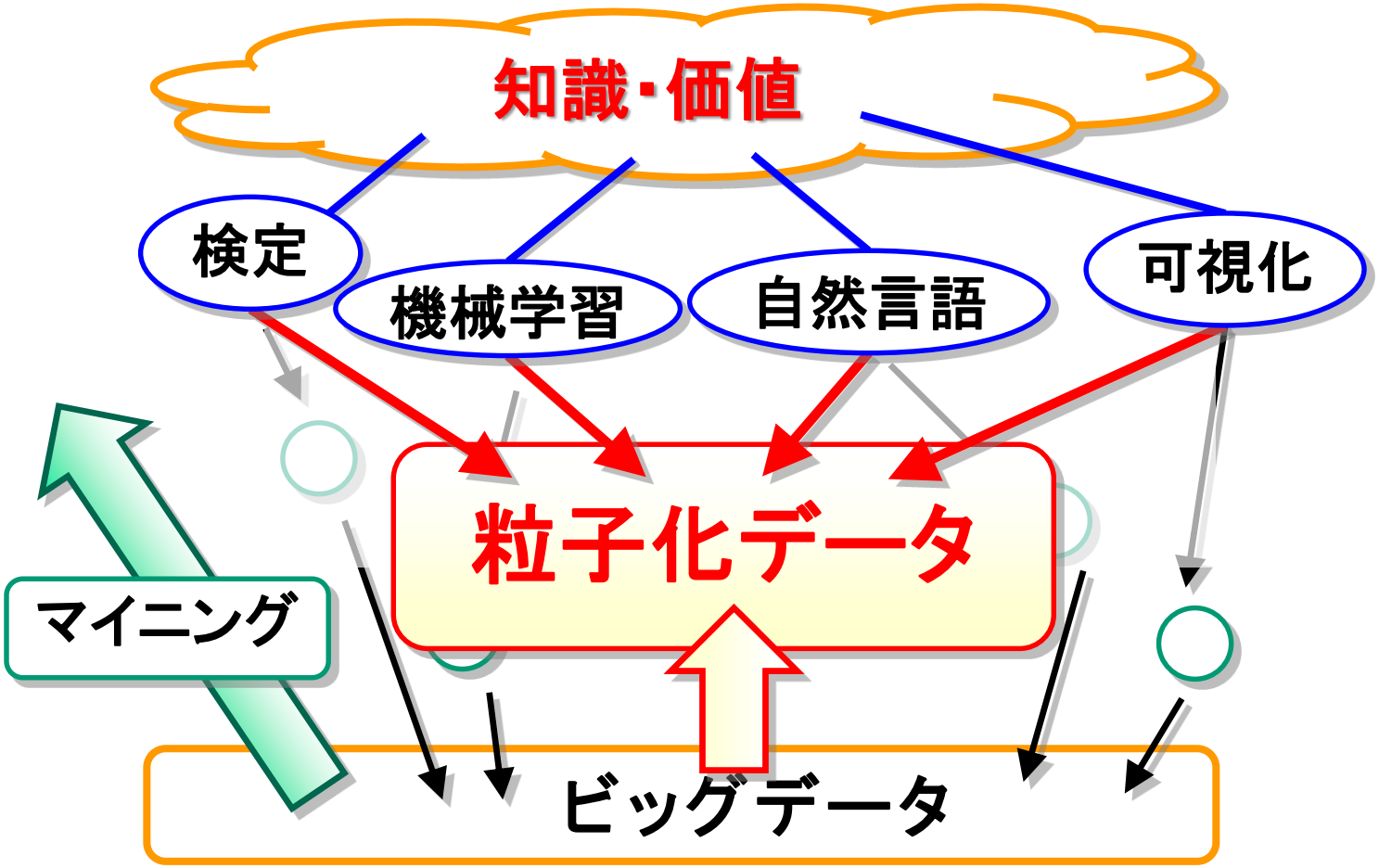


- 分類の「意味」を捉えやすそう

抽象化、現状は？

- 抽象化をすべきマイニングは、「利用しやすさ」の発想に欠ける
- 手法・問題ごとに場当たり／散発的で、「理解しやすさ」が欠落

→ 開発の容易さ、精度向上、理解しやすさ、技術の統一的な俯瞰

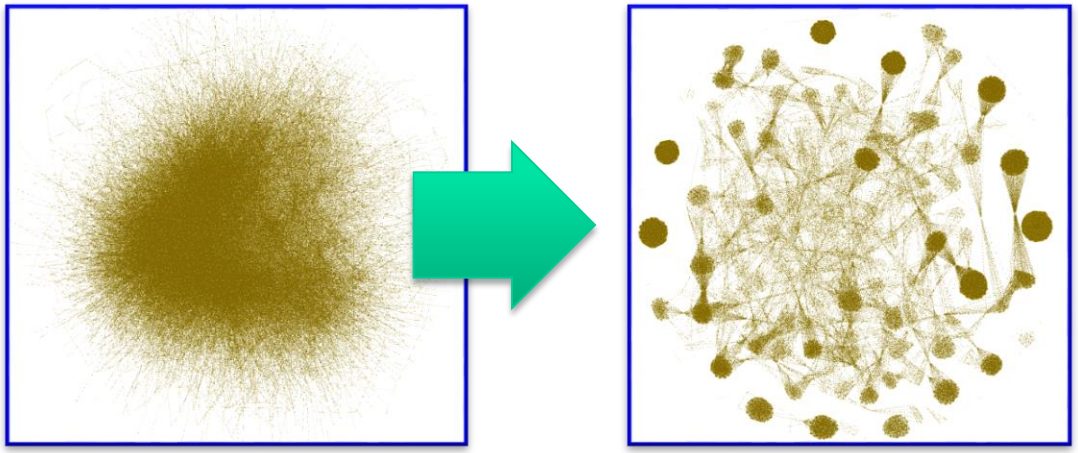


データ研磨の威力

- 帝国データバンク様、企業間取引データ

  点：企業、線：取引ある、似てる：共通友達 PMI ≧0.6

	研磨前	研磨後
頂点数	3,282	3,282
枝数	35,168	73,132
クリーク数	32,953	343



買い物データを用いた、顧客が健康志向かどうかの予測精度

	クリーク	Newman	グラフカット
研磨前	60.60%	59.70%	60.03%
研磨後	**71.36%**	**62.76%**	**67.78%**

人工生成したベンチマークデータにおける、意味的構造の検出率

	研磨	Newman	グラフカット
ノイズ10%	**100.00%**	68.74%	76.10%
ノイズ40%	**99.69%**	7.91%	77.03%

研究体制

データ研磨
アルゴリズム

宇野 毅明　情報学研究所

実応用、可視化
ビジネスデータ

羽室 行信　関西学院大学

意味構造解析
機械学習

山本 章博　京都大学

インタラクション
ユーザ解析

中小路 久美代　京都大学

実データとユーザ行動の解析から課題抽出を行い、
データ研磨アルゴリズムと粒子の持つ意味構造の解析手法を構築

誘導二部グラフの効率良い列挙

国立情報学研究所 和佐 州洋，宇野 毅明

<u>入力</u>：ラベル付き無向グラフ$G = (V, E)$

<u>出力</u>：連結二部グラフを誘導する頂点集合すべて

<u>主結果</u>：グラフの**<u>縮退性</u>**を用いて，<u>解1つあたり$\boldsymbol{O(k)}$**時間**</u>で列挙

平面グラフなどの定数縮退数のグラフに対して最適

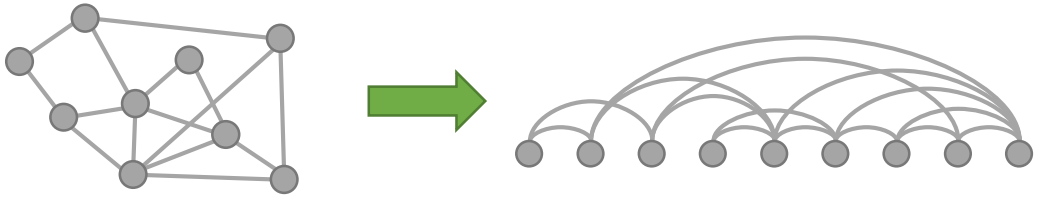


Fig. k−縮退グラフとその縮退列

k: グラフの縮退数

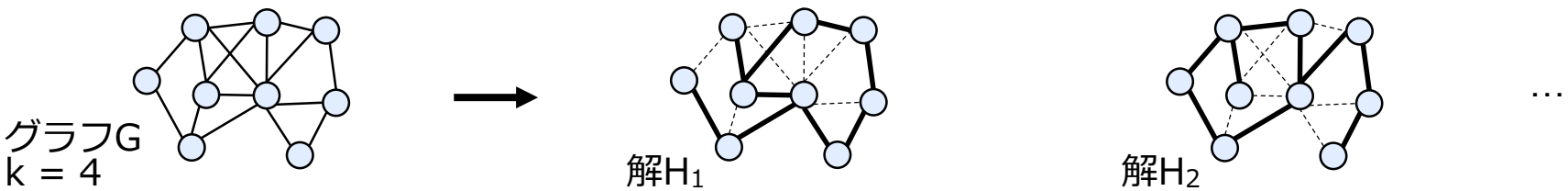
グラフ中の内周k以上の部分グラフと誘導部分グラフの効率良い列挙

栗田 和宏[1], Alessio Conte[3], 和佐 州洋[2] , 宇野 毅明[2], 有村 博紀[1] 1. 北海道大学, 2. 国立情報学研究所 3. Universit'a di Pisa

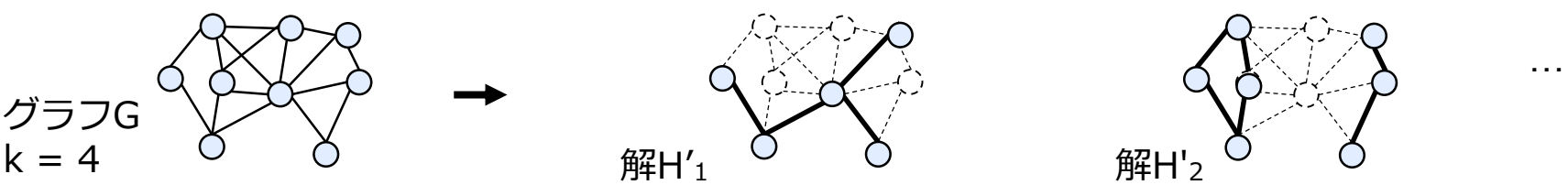
内周: グラフに含まれる最小サイクル長

入力. グラフG, 整数k
出力. G中の内周k以上の部分グラフ

<u>定理1</u>. 内周k以上の部分グラフHを1つあたり
O(|V(H)|)時間で列挙できる.



<u>定理2</u>. 内周k以上の誘導部分グラフH'を1つあたり
O(|N[V(H')]|)時間で列挙できる.



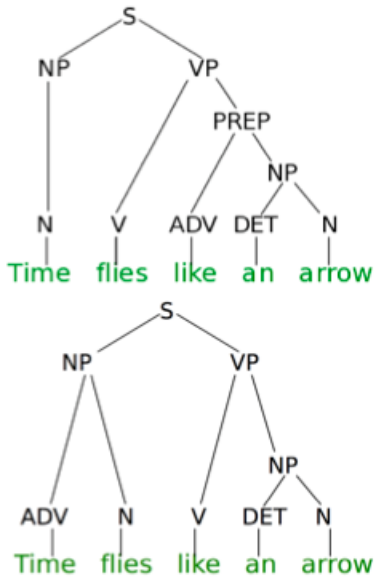
文脈自由文法による構文木の集合を表現する決定グラフの構築

網井 圭 (京都大学), 西野 正彬 (NTT CS研), 山本 章博 (京都大学)

文脈自由文法により記号列の構文木を取得可能 (構文解析)

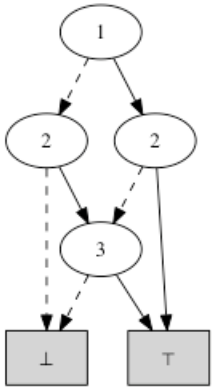
記号列に対し**全構文木**を**決定グラフ**で**圧縮表現**

構文木の集合



構築

決定グラフ



- 構文木をコンパクトに表現
- 大きさに線形な時間で演算
  - 制約条件を満たすものを抽出
  - 確率文脈自由文法に対し確率最大の構文木を探索

演算の効率は決定グラフの構築時間と大きさに依存

研究結果: 効率的な構築法を提案し大きさに上限を与えた